Interactive visualization of high-dimensional data via continuous low-dimensional projections

Mid-candidature review — 25 Feburuary, 2020

Nicholas Spyrison, B.Sc

Monash University

Faculty of Information Technology

Thesis Supervisors

Prof. Kimbal Marriott

Prof. Dianne Cook

Committee Members

Dr. Maxime Cordiel

Dr. Shirui Pan

Chair

Assoc. Prof. Bernhard Jenny

Contents

1 Introduction

1.1 Motivation

The term exploratory data analysis was coined by Tukey (1977), who leaves it as an intentionally broad term that encompasses the initial summarization and visualization of a data set. This is a critical first step of checking for realistic values and validating model assumptions. It may be tempting to review a series of summary statistics to check model assumptions. However, there are known datasets where the same summary statistics miss glaringly obvious visual patterns (Anscombe 1973; Matejka and Fitzmaurice 2017). It is strikingly simple to look at the wrong, or incomplete set of statistics needed to validate assumptions. Data visualization is fast, versatile, and robust relative to the alternative of numeric statistical summarization. Data visualization does and must remain a primary component of data analysis and model validation.

Consider tabular data containing many attributes (variables). Visualization of this ubiquitous type of dataset is key to its understanding and exploration. Visualization of spaces in more than 3 dimensions quickly becomes problematic. We will discuss the use of linear projections to mitigate this obstacle. The motivation for this research is two-fold: expand the dimensionality support and improve the understanding from such visualizations.

1.1.1 Current state of the field

Consider plotting 2 variables as an XY scatterplot. To add a 3rd variable append the z-axis orthogonal (right angle to) the XY plane. Adding a 4th dimension is not so easily solved. To resolve this, we will use scatterplot matrices to introduce axes and bases. Then we generalize to linear projections before discussing principal component analysis and introducing tours.

Scatterplot matrices (Chambers et al. 1983) plot every combination of variable pairs and views them in a matrix. This is a good method to check variable ranges and extreme values but is not the ideal visualization. Consider a bar stool as in figure **??**. Given 3 still-frames of the variable-pair (that is, a square-on view from each dimension) does not necessarily convey the full information of the data. For instance, a three-quarter perspective helps relate the square on images and are ubiquitous in assembly and machining instructions. This perspective relates information contained in the view of sides. Mathematically, we describe the angle of view as a `basis`. Bases (plural of basis) are depicted as unit axes, they point in the direction each dimension is oriented. The axes for the three-quarter perspective differ from the square-on views in that the directions are a combination of 2 and 3 variables respectively, rather than one variable mapped fully to the horizontal or vertical.

Building on pairs of variables, an arbitrary `p` variables can be projected down to 2 dimensions. This seems unintuitive at first, but we have already discussed some trivial cases with scatterplot matrices. Consider figure **??** again. The first 3 cases plot pairs variables, while the third direction extends directly beyond the XY plane. These are trivial projections of 3- down to 2- dimensions, where the 3rd variable has no contribution with a row of zeros in the basis. The three-quarter perspective is more interesting as more than one variable contributes to each direction. The resulting plot is said to be a *linear projection* when the basis is produced with an affine transformation, that is, any transformation in which all parallel lines remain parallel. One crucial aspect of linear projections is that they are interoperable back to the original attributes; any observation identified in any linear projection can be mapped back to its variable values. Recently, some non-linear projection techniques such as self-organizing maps (Kohonen 1990) and t-distributed stochastic neighbor embedding (Maaten and Hinton 2008) have received a substantial following. However, due to their non-linear transformation observations cannot be mapped back to the original attribute-space and interpretation becomes opaque. For this reason, non-linear transformations are precluded from the exploration of data- or parameter-spaces.

Principal component analysis (PCA; Pearson 1901) is one common way of identifying projections to consider. In PCA the new components are formed from a linear combination of the original attributes. The new components must be orthogonal to all previous components and ordered by descending variation explained. A pair of these new components are then viewed as an XY scatterplot. This has the added benefit of viewing the most variation in as few dimensions as possible. The scree plot/test (Cattell 1966) can be used on the components to quickly zero in on a space that has the intrinsic dimensionality of the data. This is a common data processing step once the number of attributes becomes sizable (larger than 10 or so).

1.2 From discrete to continuous

The above methods have suggested a *discrete* number of linear projections to look at. At the same time, the stool example illustrates that looking at intermediate views improves understanding. A *continuous* animation of the object being rotated would improve this understanding even further. This is analogous to the idea of a data visualization *tour*(Asimov 1985; Buja and Asimov 1986). A tour produces a relatively high number of linear projections and views them in quick succession, typically as an animation. When the bases have a relatively small change in the contributions the projections are much closer. Single points and features can be tracked and follow from projection to projection allowing for a better understanding of the local structure.

One key feature of tours is the selection of the path animated. Asimov originally purposed a *grand* tour. In the grand tour, several target planes are identified somewhat close to the starting basis. Many interim planes are found between target planes via geodesic interpolation. Figure 1 illustrates a simplification of this process. A *little* tour (Wickham et al. 2011) starts on a basis with only 2 variables contributing and trades the contribution one variable for another single variable. This effectively animates between the planes displayed in a scatterplot matrix. A *local* tour explores a dense area of the possible projections around the initiated basis. It walks a short distance away before returning to the starting basis and selecting a new random frame very nearby.

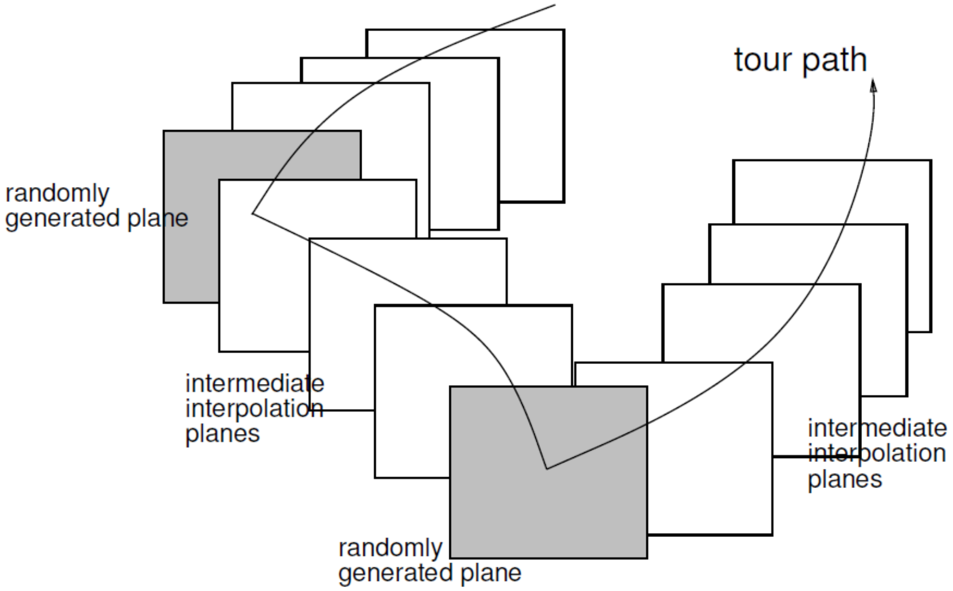

tour path
randomly generated plane
intermediate interpolation planes
intermediate interpolation planes
randomly generated plane

Figure 1: Simlified illustration of a grand tour target and interim frames (Buja et al. 2005)

The *guided* tour (Cook, Buja, and Cabrera 1993) makes use of projection pursuit (Kruskal 1969, 1972) to identify a path. Here a selected index function is selected. Starting from the initial basis several nearby bases are checked to see if they perform better on the index function. If so, this new basis is moved and nearby bases are checked again in an iterative process. If the closest bases do not perform better, bases slightly further away are checked as in simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983). If no better basis is identified within the selected stopping criterium, the algorithm stops.

The discrete methods above identity bases that highlight some feature. The above tours can help understand the structure leading to and from certain projections. To explore higher dimensional spaces visually, we need *human-in-the-loop* (Karwowski 2006) tools. We need to be able to choose and direct the change to a basis. (Cook and Buja 1997) introduces the idea of the *manual* tour. In a manual tour, an individual variable is selected and its contribution to the projection basis is then controlled. The manual tour can play a predefined manual path or be used interactively, where the user defines each subsequent step while viewing the current projection. The later, human-in-loop method we define as *User-Controlled Steering* (UCS) specifically as the interactive application of the manual tour.

1.3 Research objectives

Data and models are typically high-dimensional, with many variables and parameters. Developing new methods to visualize high dimensions has been a pursuit of statisticians, computer scientists and visualization researchers for decades. As technology evolves examining, extending, and assessing current techniques, in new environments, for new data challenges, is an important endeavor. The primary Research Objectives (RO) can then be summarized as the following.

#TODO: reword "ideal form", drop to contribution? discuss with Di tomrorrow. 1. **What is the ideal form of an interoperable application of manual tour?**
An algorithm for the manual tour is discussed in Cook and Buja (1997). The algorithm should be better supported, especially proof for the rotation matrix identified. An application of this ideally embodies two properties: interoperable with other tour applications and allow for 'next step' interaction for UCS. Manual tours, and specifically the fine control of UCS, should allow analysts to better explore the sensitivity of structure from single variable contributions on the projection. This also acts as the foundation to be used in the remaining objectives.

2. **Do 2D interactive manual tours provide benefits over alternatives?**
   User-control manipulation around identified projections of interest theoretically allows for a better understanding of the variables contributing to the structure. However, the level of abstraction and the sheer number of basis permutations is formidable. The idea that UCS will provide digestible improvements to the understanding of structure should be validated.

3. **How can the manual tour be extended to 3D?**
   Gracia et al. (2016) compares analyst tasks between 2- and 3-D scatterplots. They find modest accuracy and error improvements for distance perception and outlier identification in favor of 3D at the cost of a relatively small increase in task time. Nelson, Cook, and Cruz-Neira (1998) similarly compare a 2D scatterplot projection and the 3D scatterplot manipulation space (of the same projection). They find a slight advantage in the sphere test and a large advantage in the cluster test in favor of the 3D

manipulation space. The manual tour should be extended to accommodate a 3D projection as analogous improvements may be obtainable.

4. **Do 3D manual tours provide benefits over alternatives?**
   #TODO: add here, can we use the same data? same app? same idea as RO2?

1.4 Methodology

This research is interdisciplinary; touring was developed by statisticians to explore physics data. Modern advances in hardware from information technology allow for 3D rending in higher quality and immersion than previously possible.

The research corresponding with RO #1 entails *algorithm design* following and further clarifying the work done in Cook and Buja (1997). In the application of the manual tour, we clarified the creation of the rotation matrix. The key to the matrix is to specify the 2 axes of rotation for the manipulation space and apply Rodrigues' rotation formula (Rodrigues 1840). In the application, attention was given to both pre-compiled tour and human-in-the-loop UCS. We provide an open-source version of the manual tour in the R package, `spinifex`, which has since been published on CRAN. This forms the foundation for future work in the remaining objectives.

For RO #2 is a controlled *experimental study* to explore the efficacy of interactive UCS compared with the benchmarks factors of PCA and the grand tour. This is designed as a within-participant study where each participant performs all factors. the study is balanced by assigning participants into one of 3 groups where the factor order is controlled by a Latin square while simulation order remains the same. The details are discussed in finer detail in section (#sec:expStudy), below.

The research for RO #3 involves *algorithm design* extending the current 2D manual tour into a 3D project. For a 2D projection, the axes basis is rotated through a 3D 'manipulation space'. In a 3D projection, such a space requires 4 dimensions. Theoretically, after the addition of a new angle of rotation, the rotation matrix must be extended to accommodate a new dimension and angle parameter. This also means that analysts have another parameter to define, further increasing their already sizable input-volume.

2 Progress since confirmation

During the candidature confirmation review (27 March 2019) we discussed exploratory data analysis, visualization of high dimensional spaces, covered the literature for tours and 3D rendering for information perception. We concluded with a process for a manual tour that allows for user-controlled steering. The appending the document was a mostly complete R package and respective paper providing an open-source application as well as clarifying the rotation matrix that was outlined in Cook and Buja (1997).

2.1 Publication

The paper has since been accepted in the R Journal and it currently undergoing editorial review. This will be published in the first issue of 2020 and available at journal.r-project.org.

2.2 Software

The R package, `spinifex (v0.1.0)` (Spyrison and Cook 2019), has been approved and hosted for public use on the Comprehensive R Archival Network, CRAN (cran.r-project.org/web/packages/spinifex/)(https://cran.r-project.org/web/packages/spinifex/index.html).

New functionality has been added to the development branch, specifically around the interactive use. A user application is developed with `shiny`(Chang et al. 2018). It allows users to explore their data without the need for coding familiarity. It features interactive or precompiled manual touring. It also contains a gallery for flagging bases which can then further be reviewed or saved as .gif and .png files.

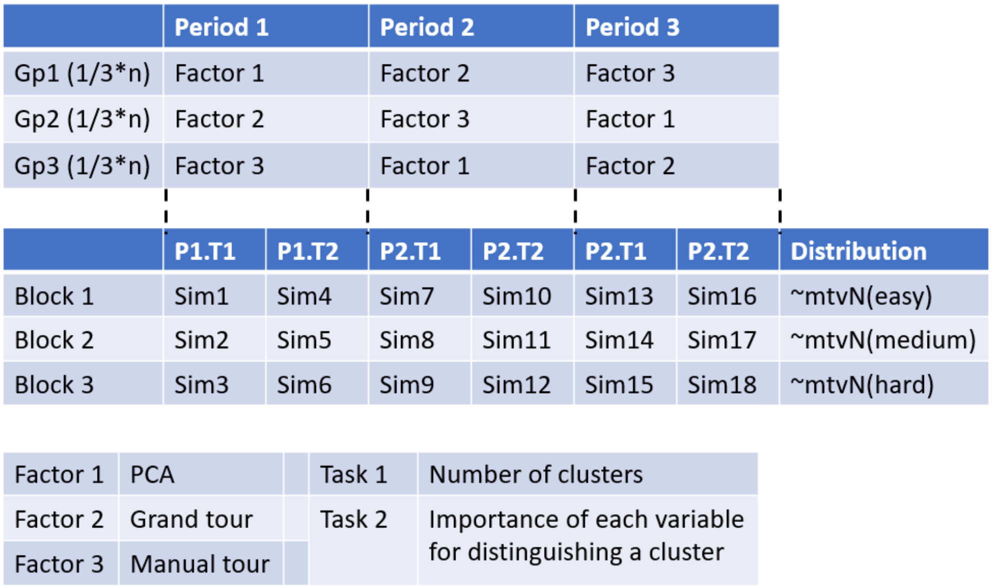
2.3 Experimental study

The prominent appeal of the manual tour is that it allows users to control the contributions of individual variables. In theory, this should enable the finer exploration of features of interest. The hypothesis we will study is *Does the finer control afforded by the manual tour improve the ability of the analyst to understand the importance of variables contributing to the structure?*

We list out an abbreviated summary of the experimental design in the subsections below. A more detailed draft of the working paper is included as an appendix.

2.3.1 Groups

Each participant will be randomly split into one of three even groups. The group controls the order of the factors that the participant was evaluated in for a latin square of the 3 factors. For instance, the order of the first group was PCA, grand, manual. Group level only impacts the order the factors are displayed while task, block, and simulation order will remain the same.



	Period 1	Period 2	Period 3
Gp1 (1/3*n)	Factor 1	Factor 2	Factor 3
Gp2 (1/3*n)	Factor 2	Factor 3	Factor 1
Gp3 (1/3*n)	Factor 3	Factor 1	Factor 2

	P1.T1	P1.T2	P2.T1	P2.T2	P2.T1	P2.T2	Distribution
Block 1	Sim1	Sim4	Sim7	Sim10	Sim13	Sim16	~mtvN(easy)
Block 2	Sim2	Sim5	Sim8	Sim11	Sim14	Sim17	~mtvN(medium)
Block 3	Sim3	Sim6	Sim9	Sim12	Sim15	Sim18	~mtvN(hard)

Factor 1	PCA		Task 1	Number of clusters
Factor 2	Grand tour		Task 2	Importance of each variable for distinguishing a cluster
Factor 3	Manual tour			

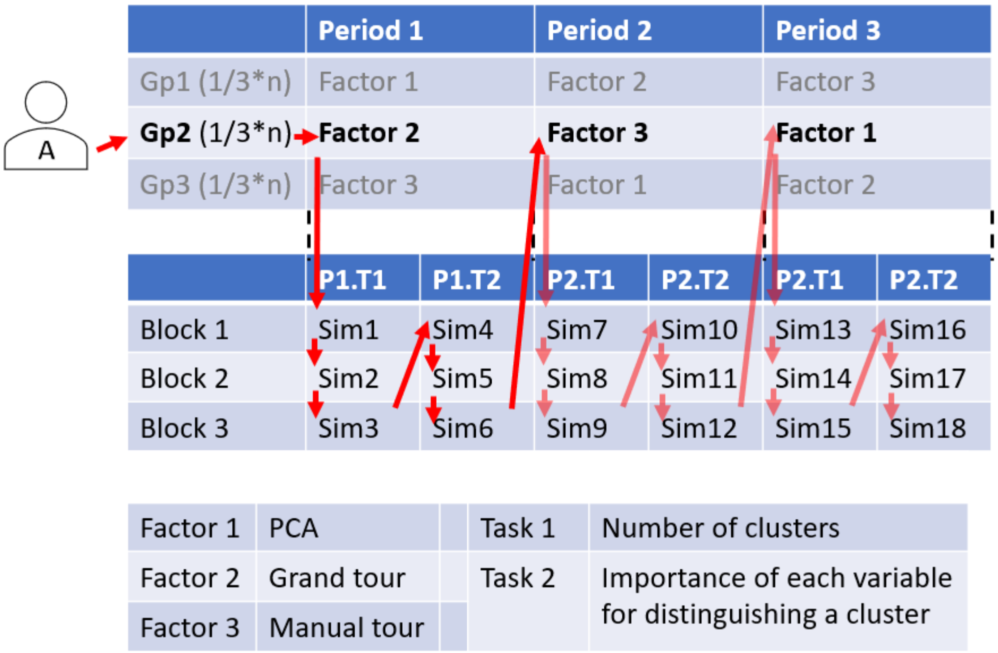
Figure 2: Experimental design setup. Participants are assigned to one of 3 even groups controlling the factor order. Within each factor, users perform 3 difficulty blocks of task 1 and then of task 2 before proceeding to the next factor. Simulations are used in a fixed order (while factor order changes). Simulations for the first block difficulty are unique samples drawn from the same distribution. Similarly, the second and third block difficulties are drawn from increasing complex distributions.

2.3.2 Factors

We will explore performance across three factors. The first factor is Principal Component Analysis (PCA). The second factor is an animated walk of interpolation frames between target bases, called a *grand* tour. The third factor allows for the manual control of the individual variable's contribution to the projection, performing a *manual* tour.

All factors are shown as a scatterplot. The basis axes projection was also illustrated to the left of the plot. They are shown in a unit circle and show the magnitude and direction each variable contributes to the projection.

The user interface was kept the same whenever possible, but the control inputs do change slightly to accommodate the differences between factors. The inputs for PCA select a pair of principal components to display on the x- and y-axes. The manual tour had the same axes selection, with the addition of a drop-down bar and slider control selecting the manipulation variable and magnitude. The grand tour comes precompiled as an animation of a 15 second showing 90 frames at 6 frames per second. The user can control the location or play/pause the animation at will.



	Period 1	Period 2	Period 3
Gp1 (1/3*n)	Factor 1	Factor 2	Factor 3
Gp2 (1/3*n)	**Factor 2**	**Factor 3**	**Factor 1**
Gp3 (1/3*n)	Factor 3	Factor 1	Factor 2

	P1.T1	P1.T2	P2.T1	P2.T2	P2.T1	P2.T2
Block 1	Sim1	Sim4	Sim7	Sim10	Sim13	Sim16
Block 2	Sim2	Sim5	Sim8	Sim11	Sim14	Sim17
Block 3	Sim3	Sim6	Sim9	Sim12	Sim15	Sim18

Factor 1	PCA		Task 1	Number of clusters
Factor 2	Grand tour		Task 2	Importance of each variable for distinguishing a cluster
Factor 3	Manual tour			

Figure 3: Example case. Person ’A’ is assigned to group 2, where they will use factor 2 (grand tour) for the first period. They perform 3 block difficulties of task 1 on simulations of increasing difficulty. Then 3 block difficulties of task 2 on unique simulations sampled from the same distributions of increasing difficulty. After this, they proceed to period 2, where they are use factor 3 (manual tour) to perform 3 block difficulties of each task. Lastly, in the third period, they use factor 1 (PCA) to perform the tasks.

2.3.3 Tasks

Within each factor, participants will perform 2 tasks. The first task asks participants to identify the number of clusters present in unsupervised data. This task served as a standard for assessing the general aptitude for this sort of high dimensional analysis as it was simpler. In application, linear discriminant analysis (Fisher 1936) or penalized discriminant analysis (Hastie, Buja, and Tibshirani 1995) are better suited for classifying such unsupervised data.

The second task is focused on the hypothesis of the study, it asks participants to identify any/all variables that were very important and somewhat important for distinguishing a given cluster from the others. For instance, which variables are very- and somewhat- important for distinguishing clusters ‘A’ and ‘B’.

2.3.4 Block difficulty

Participants will be randomly assigned to 1 of 3 even groups. Each group has a different factor order containing all factors. Both tasks will be performed in the same order. Each task will have 3 repetitions performed on new simulations that were drawn from 3 parameterizations in increasing difficulty. Each participant will go through the simulations in the same order, while their factor order will vary. Fixing block difficulty order while varying factors should mitigate potential learning bias.

3 Proposed thesis structure

3.1 Thesis structure

- Introduction – 60%
- Literature review – 80%
- Manual tour and user-controlled steering – 90%
- Experimental study – 60%
- The extension of the manual tour to 3D – 5%
- Conclusion and future plans – 0%

Proposed research timeline

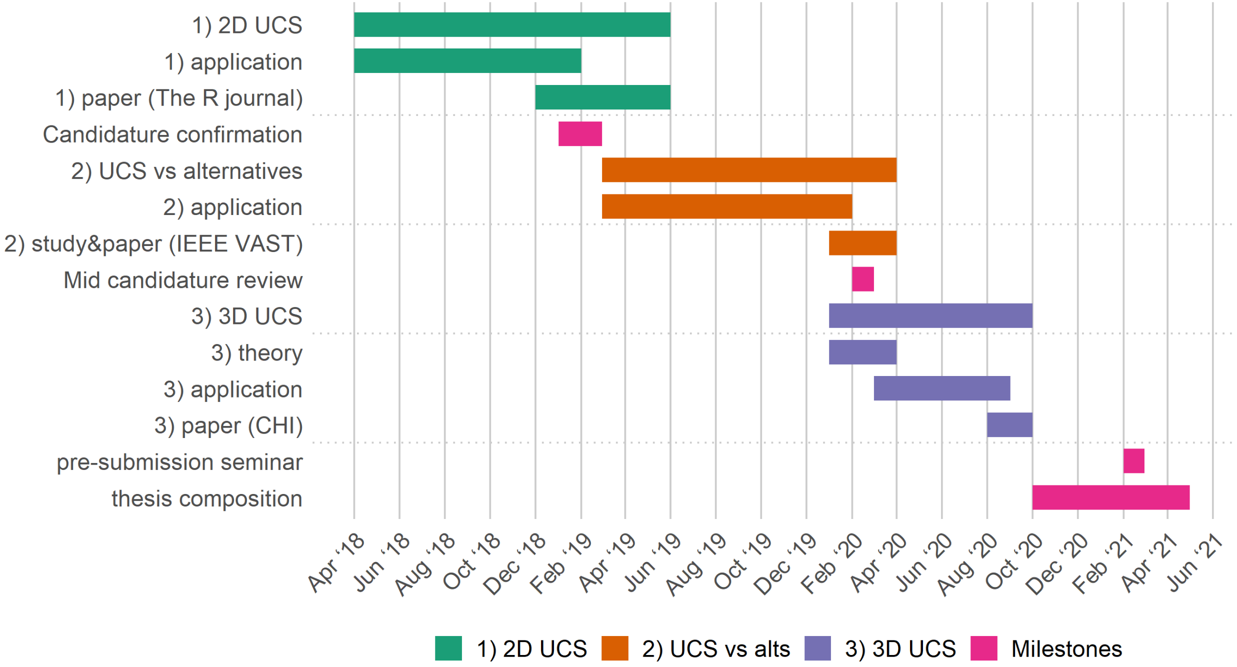


Figure 4: Proposed research timeline.

3.2 Program requirements

- WES Academic record
  - FIT5144: 2019 S1+2, **In progress**, extended to the pre-submission seminar with the unit coordinator for the usual 2 opportunities to complete.
    - Hours: 147>120 hours **Tracked**, missing the following requirments (12 hr total)
    - *Needed:* CYR 2 (A & B) – 2x 3hr
    - *Needed:* Faculty of IT Workshop 1 and 3 on Ethical Research and Publishing – 2x 3hr
  - FIT5113: 2018 S2, **Exemption**
  - FIT6021: 2018 S2, **Completed** with distinction
- myDevelopment - IT: Monash Doctoral Program - Compulsory Module
  - Monash graduate research student induction: **Completed**
  - Research Integrity - Choose the Option most relevant: **Completed**
  - Faculty Induction: **Completed**

4 Potential issues for panel to consider

4.1 Funding for human subjects

- Beverage voucher: $5 x 24 people (est) = $120

4.2 Support for conference travel

Conferences:

CHI 2021: May 8-13, 2021 Yokohama, Japan

submission: Thursday Sep. 10, 2020

https://chi2021.acm.org/

IEEE VAST - VAST 2020: 25-30 October 2020 Salt Lake City, Utah, USA

submission: Saturday, March 21, 2020

http://ieeevis.org/year/2020/info/call-participation/vast-paper-types

5 Acknowledgements

This report was created in R (R Core Team 2019), using bookdown (Xie 2016) and rmarkdown (Xie, Allaire, and Grolemund 2018).

For version control, transparency, and reproducibility, the source files are made available found at github.com/nspyrison/mid_canidature.

References

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. https://doi.org/10.2307/2682899.

Asimov, Daniel. 1985. "The Grand Tour: A Tool for Viewing Multidimensional Data." *SIAM Journal on Scientific and Statistical Computing* 6 (1): 128–43. https://doi.org/https://doi.org/10.1137/0906011.

Buja, Andreas, and Daniel Asimov. 1986. "Grand Tour Methods: An Outline." In *Proceedings of the Seventeenth Symposium on the Interface of Computer Sciences and Statistics on Computer Science and Statistics*, 63–67. New York, NY, USA: Elsevier North-Holland, Inc. http://dl.acm.org/citation.cfm?id=26036.26046.

Cattell, Raymond B. 1966. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research* 1 (2): 245–76.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. "Graphical Methods for Data Analysis."

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, and Jonathan McPherson. 2018. *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny.

Cook, Dianne, and Andreas Buja. 1997. "Manual Controls for High-Dimensional Data Projections." *Journal of Computational and Graphical Statistics* 6 (4): 464–80. https://doi.org/10.2307/1390747.

Cook, Dianne, Andreas Buja, and Javier Cabrera. 1993. "Projection Pursuit Indexes Based on Orthonormal Function Expansions." *Journal of Computational and Graphical Statistics* 2 (3): 225–50. https://doi.org/10.1080/10618600.1993.10474610.

Fisher, Ronald A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2): 179–88. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Gracia, Antonio, Santiago González, Víctor Robles, Ernestina Menasalvas, and Tatiana von Landesberger. 2016. "New Insights into the Suitability of the Third Dimension for Visualizing Multivariate/Multidimensional Data: A Study Based on Loss of Quality Quantification." *Information Visualization* 15 (1): 3–30. https://doi.org/10.1177/1473871614556393.

Hastie, Trevor, Andreas Buja, and Robert Tibshirani. 1995. "Penalized Discriminant Analysis." *The Annals of Statistics*, 73–102.

Karwowski, Waldemar. 2006. *International Encyclopedia of Ergonomics and Human Factors, -3 Volume Set*. CRC Press.

Kirkpatrick, Scott, C. Daniel Gelatt, and Mario P. Vecchi. 1983. "Optimization by Simulated Annealing." *Science* 220 (4598): 671–80. https://doi.org/10.1126/science.220.4598.671.

Kohonen, Teuvo. 1990. "The Self-Organizing Map." *Proceedings of the IEEE* 78 (9): 1464–80.

Kruskal, Joseph B. 1969. "Toward a Practical Method Which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation Which Optimizes a New 'Index of Condensation'." In *Statistical Computation*, edited by Roy C. Milton and John A. Nelder, 427–40. Academic Press. https://doi.org/10.1016/B978-0-12-498150-8.50024-0.

———. 1972. "Linear Transformation of Multivariate Data to Reveal Clustering." *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences* 1: 181–91.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9 (Nov): 2579–2605.

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 1290–4. Denver, Colorado, USA: ACM Press. https://doi.org/10.1145/3025453.3025912.

Nelson, Laura, Dianne Cook, and Carolina Cruz-Neira. 1998. "XGobi Vs the C2: Results of an Experiment Comparing Data Visualization in a 3-d Immer- Sive Virtual Reality Environment with a 2-d Workstation Display." *Computational Statistics* 14 (1): 39–52.

Pearson, Karl. 1901. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rodrigues, Olinde. 1840. *Des Lois Géométriques Qui Régissent Les Déplacements d'un Système Solide Dans L'espace: Et de La Variation Des Cordonnées Provenant de Ces Déplacements Considérés Indépendamment Des Causes Qui Peuvent Les Produire.*

Spyrison, Nicholas S., and Dianne Cook. 2019. *Spinifex: Manual Tours, Manual Control of Dynamic Projections of Numeric Multivariate Data* (version 0.1.0.9000). https://github.com/nspyrison/spinifex/.

Tukey, John W. 1977. *Exploratory Data Analysis.* Vol. 32. Pearson.

Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2011. "Tourr: An R Package for Exploring Multivariate Data with Projections." *Journal of Statistical Software* 40 (2). https://doi.org/10.18637/jss.v040.i02.

Xie, Yihui. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown.* Boca Raton, Florida: Chapman; Hall/CRC. https://github.com/rstudio/bookdown.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide.* Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.